DOCUMENT RESUME

ED 294 927 TM 011 694

AUTHOR Schoon, Craig G.; And Others
TITLE A Critique of Difficulty Estimation Methodologies in the Setting of Cut Points and a Discussion of an Alternative Methodology: The Direct Standard Setting Method.
PUB DATE Apr 88
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cutting Scores; Difficulty Level; Estimation (Mathematics); *Item Analysis; *Licensing Examinations (Professions); Test Construction; Test Items; *Test Validity
IDENTIFIERS *Direct Standard Setting Method

ABSTRACT

The determination of appropriate cut scores is a critical step in the development of licensing and certification examinations. Passing point methodologies based on the estimation of item difficulties are underlain by the estimation of the probability of a correct response to items by a hypothetically minimally competent candidate. The Angoff technique is typical of such methods. The use of such constructs as "minimally competent" and "probability of a correct response," however, in setting passing points is problematical. Little evidence exists concerning the utility of these constructs, and it is difficult to provide them with operational definitions or validate them against external criteria. As an alternative to these methods, a Direct Standard Setting Method has been developed upon the following bases: (1) emphasis on judges' estimation of required level of candidates' knowledge rather than probability of correct response; (2) development of a standard that provides the definition of the performance necessary for examinees to be certified; (3) an understanding by the judges that their ratings of questions will be translated into performance standards; and (4) a methodology that is easy to follow for judges from any profession. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED294927

A CRITIQUE OF DIFFICULTY ESTIMATION METHODOLOGIES IN THE SETTING OF CUT POINTS

AND

A DISCUSSION OF AN ALTERNATIVE METHODOLOGY: THE DIRECT STANDARD SETTING METHOD

by

Craig G. Schoon, Gerald Rosen, J. Patrick Jones

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Professional Examination Service
475 Riverside Drive
New York, New York 10115

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

CRAIG G. SCHOON; GERALD A. ROSEN; J. PATRICK JONES

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Annual Meeting of the
American Educational Research Association
New Orleans, Louisiana
April, 1988

TM 011 694

The determination of appropriate passing points is a critical step in the development of licensing and certification examinations. Based on this step, passing points are derived that will be used to make licensing decisions that determine who will be allowed to earn a living through the practice of the profession, and that further carry the implication that those selected have the knowledge, ability, and skill to practice the profession safely and without harm to the public. The decisions made from certification examinations, although less critical than licensing decisions, carry considerable weight in that the successful candidate can use titles conferred through the certification along with the benefits and rights that the credential implies.

The important implications that derive from credentialing decisions mandate that we who are involved in the selection and use of methodologies and procedures to establish passing standards, scrutinize these methodologies and procedures for their logicality, validity, and reliability.

The intent of this paper is, first, to present a critique of passing point methodologies based on estimation of item difficulties, and then to propose an alternative passing point methodology called the Direct Standard Setting Method (DSSM)

The item difficulty estimation methodologies used to set passing points for credentialing examinations are based on the estimation of the probability of a correct response to items by a hypothetically minimally competent candidate. Thus, for example, the Angoff technique (Angoff, 1971) asks judges to estimate, for each item on an examination, the probability that a minimally competent candidate will answer the item correctly. The passing point is then computed by averaging the ratings of all judges across all items, the average rating being set as the passing point. We will begin our critique by examining the assumptions upon which such a methodology is based.

The Angoff technique assumes that judges are familiar with the characteristics of minimally competent candidates in their professions and that they can estimate, based on this

familiarity, the probability that such candidates will answer given items correctly. The Angoff method is valid to the extent that judges are indeed familiar with the characteristics of minimally competent candidates, and can indeed estimate with accuracy the probability that they can answer items correctly. Distressingly little evidence exists however, to support these assumptions. Empirical evidence regarding these assumptions could be provided through validation studies where there was a reliable independent measure of competency status, and where the point of minimal competence on an examination could be identified. Candidates thus identified as minimally competent should then be accurately selected using the passing point of the examination in question. Indirect evidence on this question, however, was collected by Schoon, Gullion, and Ferrara (1979). They had expert judges in the field of Occupational Therapy set a passing point using the Angoff technique. The examination had been administered to a sample of currently practicing occupational therapists. The passing point set by the judges would have failed over one third of the practicing occupational therapists who took the examination. Other studies have also found that application of the Angoff and similar passing point methodologies fail a high percentage of individuals already identified as competent (Glass, 1978).

The null hypothesis would lead one to assume that judges have difficulty envisioning the characteristics of minimally competent candidates. One might also question, further, if judges can accurately estimate the probability that such candidates will respond correctly to given multiple choice questions. The authors have found that judges in the credentialing fields do have difficulty in envisioning the characteristics of minimally competent candidates in their professions, and further, that they have difficulty in estimating probabilities of correct responses to individual items.

In psychological measurement, a construct is useful to the extent that it can be reliably measured on the basis of its operational definition. For the construct of "minimal competence" to be useful in the setting of passing points, we must be able to operationally define

it for judges. The problems that judges have in using this construct to set passing points most probably results from the difficulties encountered in its operational definition. This is especially true in the case of licensing and certification examinations, where the behaviors of interest are samples from a large and complex universe of cognitive, motor, and psychomotor skills. The construct of minimal competence is applicable and useful where the behaviors to be measured can be precisely defined, e.g., the multiplication of 3-digit numbers, and where the measurement instrument corresponds to the behavior to be measured, i.e., multiplication skills. The construct is less useful the more complex the behavior domain to which it refers becomes, and the more difficult it becomes to operationally define it. Thus, for most occupations, the construct of minimal competence is so complex that its operational definition is problematical. The construct of minimal competence, then, has limited usefulness as an aid to setting passing points in the credentialing fields because it is difficult to operationally define in terms of the many and complex cognitive, motor, and psychomotor skills that compose competent behavior. Furthermore, the measurement tools used to measure behavior, multiple choice examinations, are far removed from the actual behaviors of the profession, so that it becomes difficult to estimate how minimally competent individuals will respond. Given the difficulty in operationally defining the construct of minimal competence, and in having judges use this definition for accurate estimations of behavior, its use for the purpose of setting passing points on credentialing examinations is problematical.

Let us examine another characteristic of item difficulty estimation methods. The question asks judges to estimate the probability that minimally competent candidates will respond correctly to given questions. The Angoff question, then, is clearly asking judges to estimate the difficulty of given questions for the minimally competent. The process of setting passing points on credentialing examinations, however, is one of setting performance standards. One may question, therefore, whether estimating the difficulty of an examination is what one wants to do if one's task is to set a performance standard for a licensing or certification examination. As the Angoff question is used in most applications, judges are asked to make their

difficulty estimates without taking factors such as item importance or item criticality into account. That is, Angoff estimates can be done virtually independently of the factors upon which a performance standard should be based, such as item importance and criticality. The Angoff estimate concerns only the estimated difficulty of given items to hypothetically minimally competent candidates. A more appropriate procedure for setting standards would be to have judges estimate how candidates **should** perform on items given the content of the item, its importance, its criticality, and its difficulty.

If one takes the Angoff methodology at face value, there are other difficulties in the interpretation of passing points yielded from application of the procedure. For example, the passing point is set using the Angoff technique at the average rating of judges over all items, and this average can be interpreted as the estimated percentage of minimally competent candidates that will pass the examination. If one assumes, then, that judges are accurately estimating the percentage of the minimally competent that will pass the items on the examination, then the passing point derived from their ratings should indicate the percentage of minimally competent candidates that will pass the exam. Thus, if the passing point was estimated to be 70%, then 70% of the minimally competent should pass, and 30% of the minimally competent should fail. If the passing point is set at 60%, then 40% of the minimally competent would fail the exam. A face value interpretation of Angoff results, then, indicates that the selected passing point should fail (1 - Angoff passing point) minimally competetent candidates. If the Angoff methodology is being used to set the passing point where minimally competent candidates will be selected, then one should account for the estimated percentage of the minimally competent that the methodology logically indicates will be failing the examination. The authors are not aware of any studies that have taken account of the logical implication of this aspect of the Angoff technique.

Another logical implication of a face-value application of the Angoff technique is as follows: high Angoff ratings, which indicate that a high percentage of the minimally competent

will pass the items, result in stringent passing points, which in practical applications tend to fail many candidates. Thus, if Angoff ratings resulted in a passing point of 80% (i.e., 80% of the minimally competent will pass these items), the practical result is a very difficult examination that is difficult to pass even for the most competent. Conversely, low Angoff ratings, indicating that a low percentage of the minimally competent will pass the examination, usually results in examinations that pass high percentages of candidates. Here, imagine an Angoff-generated passing point of 60%, indicating that the judges estimate that the minimally competent would find the examination difficult, would result in an examination with a passing point of 60% of the items. In short, there is an inverse relationship between Angoff ratings and the difficulty of resulting passing points. Angoff ratings indicating an easy examination often result in very difficult examinations, and Angoff ratings indicating a hard examination often result in relatively easy examinations.

Let us extend our logical analysis of the Angoff methodology, finally, to a consideration of how the technique applies to the purpose for which passing points are set on credentialing examinations. Passing points are set on credentialing examinations in order to determine the performance standard one must attain to be classified as at least minimally competent, in the case of licensing examinations, or as having attained a level of achievement necessary to receive certification. The Angoff methodology yields such standards by assuming that judges know the performance levels of the minimally competent, and translate this knowledge into the performance standard. This standard is then used to define as minimally competent those who achieve the standard. The assumption of the Angoff methodology, then, is that one is simply applying a pre-existing knowledge of the desired performance into a standard, which then serves to define those who achieve it as minimally competent, etc. It is the opinion of the authors that this assumption has little in the way of empirical evidence to support it, and also that it places a burden on judges to have a familiarity with minimally competent performance that few if any actually possess.

Difficulties such as those described above make it difficult for judges to make their ratings. They often do not clearly understand what it is that they are being asked to do. And, not only do judges find the Angoff technique difficult to apply, they also may not clearly understand how their ratings will eventuate in producing a passing standard. That is, many judges do not clearly understand that an Angoff difficulty rating translates directly into a passing point standard: i.e., if your Angoff ratings for a test average out to 70%, then the passing point is set to 70%. Because the setting of passing points in the credentialing area has such critical implications, judges whose ratings are used to set them should have a thorough understanding of exactly how their ratings translate into a passing point standard. This is particularly true if such an understanding would have a significant impact on one's ratings. The item difficulty estimation procedure can be criticized, then, because the instructions are difficult to apply, and also because the manner in which the results are applied can be confusing to the judges. In short, it would be desirable to have a passing point methodology whose instructions are clear, easily understood and easily applied. It is also desirable that the mannerin which the procedure produces a passing point be clearly perceived by judges, since this perception may well influence judges' ratings.

In summary, our analysis of item difficulty estimation techniques indicated that the use of such constructs as "minimally competent," and "probability of correct response," in the setting of passing points is problematical. Little evidence exists as to the utility of these constructs; it is difficult to provide them with operational definitions, and the opportunity to validate them against an external criterion is rare. We may rightly, then, be wary of the use of terms such as "minimally competent" in methodologies used to establish passing points for credentialing examinations. The use of such constructs in instructions to judges may mislead them into thinking that their ratings have more meaning than they actually have, and the constructs may also misleadingly provide a patina of respectability to the methodology that is inherent in the face validity of the constructs rather than in the methodology itself. The above critique also indicates that the Angoff technique suggests logical outcomes that are

problematical: passing points derived with the technique suggest that a percentage of the minimally competent will fail, and high Angoff ratings, indicating an easy examination for the minimally competent, often eventuate in very stringent passing points that fail the majority of candidates. Our analysis suggests that an alternative approach to setting passing points be formulated.

Shepard (1980), in her review of the various passing point methodologies, suggested that an alternative approach to the Angoff, Nedelsky, and Ebel techniques would be to "... imagine sets of items of similar difficulty and then state what proportion of such items should be answered correctly." [p. 176] She points out that such an alternative approach borrows from both the Ebel (1979) and the Jaeger (1978) methods.

Following Shepard's suggestion, the authors have formulated a passing point methodology called the Direct Standard Setting Method. This new methodology is based on the following assumptions:

1. It does not use constructs such as "minimally competent" and "probability of correct response;"

2. If judges are familiar with the content of an examination, they will be able to estimate how much of that content a candidate should know in order to receive the credential in question more easily than they can estimate the probability of correct response of candidates of predefined ability to the content;

3. The method produces a standard that then provides the definition of the performance necessary to obtain the credential. The emphasis of the technique is therefore on judges' familiarity with the content of an examination, rather than on the response characteristics of hypothetically minimally competent candidates;

4. A standard setting question must ask judges to estimate how much a candidate should know rather than how a candidate will respond; thus, the method is truly a standard setting method, not an exercise in the estimate of the examination difficulty.

5. The purpose of a standard setting question should be clear and unequivocal to judges, i.e., they should be aware that their ratings will be directly translated into a performance standard that will serve as the definition for performance required to receive the credential in question.

6. The method should be easy to use for judges from any profession, and not dependent on judge sophistication in anything other than the content of their profession.

The Direct Standard Setting Method (DSSM) can be applied by having judges rate items one at a time, as with the Angoff method, or by having them rate several items at a time. The single-item question for judges is as follows:

> Given a large sample of questions that have the same content, importance, criticality, and difficulty as the one you are rating, what percentage of these questions should a candidate answer correctly in order to receive the credential conferred with this examination.

The judge is setting a performance standard on each question, treating each question as a sample from a universe of questions of the same content, importance, criticality and difficulty. A performance standard would be derived by summing and averaging judges' ratings over all questions.

The DSSM can be applied, as Shepard (1980) suggested, to groups of questions. Thus, for example, judges can be presented with a group of ten questions and asked to rate them as follows:

> How many of these ten questions should a candidate answer correctly in order to receive the credential conferred by this examination? Please take into consideration not only the content of these questions, but also their importance, criticality, and difficulty. Your ratings will translate directly into the passing point for the examination.

In order to make their DSSM ratings, judges should receive instruction and engage in discussion regarding the process of setting standards based on familiarity with the content of the profession. The typical instruction for judges for using the Angofff technique involves having them attempt to describe and visualize attributes of the minimally competent individual in their profession. Judges then attempt to translate this conceptualization of a minimally competent candidate into probabilities of correct response to individual questions. Discussion prior to making DSSM judgments, on the other hand, should deal with how judges should make estimates of performance standards based on their familiarity with the content of the profession and their knowledge of required performance standards for receiving the credential. The process of making DSSM ratings is one of setting performance standards;

the process of making Angoff ratings is one of estimating item difficulties. This distinction should be explained to judges and discussed with them. Setting a performance standard for a given item will require the judge to take into account the item's difficulty, importance, and criticality. The DSSM rating expresses judges' estimates of how one **should** perform on a given item or set of items given the item's difficulty, importance, and criticality in relation to performance standards deemed necessary for receiving the credential in question.

The DSSM, as Shepard (1980) suggested, borrows from both the Ebel (1979) and Jaeger (1978) methods. It borrows from the Ebel method in taking item importance, or criticality, into account in setting standards. It follows Jaeger (1978) in making the question a standard setting question rather than an item difficulty estimation question. The DSSM differs from the Ebel (1979) method, on the other hand, in that it requires judges to make their ratings directly in response to a given item or set of items, rather than to fill in the cells of an item importance by item difficulty matrix. The DSSM differs from the Jaeger (1978) method in that it allows judges to make a percentage correct estimation rather than the absolute yes/no decision required by Jaeger's approach.

Let us take as an example the DSSM judgment of items for a licensing examination. The purpose of licensing is to protect the public against harm through the practice of unqualified practitioners. Items on licensing examinations, accordingly, should be related to entry level non-trivial information. A performance standard should be established for items on licensing examinations that is directly related to each item's importance and criticality for successful entry practice. For example, given two items for a licensing examination in psychology, one relating to the calculation of the standard error of measurement, and the other to recognition of the symptoms of depression, one's DSSM ratings would probably assign a higher performance standard to the latter rather than the former question, because knowledge of the signs of depression would be more important for protecting the public than knowledge of calculating a standard error of measurement.

Each of the two questions given as examples above, however, could vary in difficulty depending on the type of question asked and on many other factors. It would be expected that one would also take difficulty into account in making a DSSM rating, giving a higher performance standard to difficult questions and a lower performance standard to easy questions. The most critical factor in making a DSSM rating, however, would be the importance and criticality of the item. Again, this procedure differs from Angoff ratings, which only take into account the difficulty of items.

DSSM instructions, then, should lead judges through a discussion where ratings are given to items in several content categories, and where item difficulty varies within a given content category. Judges should make DSSM ratings on, say, 10 items that differ in content, and should then discuss their ratings, taking note of interjudge rating variations. The use of commonly accepted procedures for minimizing interjudge variation is recommended.

BIBLIOGRAPHY

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), **Educational and Psychological Measurement**, 1971, 36, 35-50.

Ebel, R. L. **Essentials of educational measurement** (3rd ed.), Englewood Cliffs, N.J.: Prentice Hall, 1979.

Glass, G. V. Minimum competence and incompetence in Florida. **Phi Delta Kappan**, 1978, 59, 602-605. (a)

Jaeger, R. M. **A proposal for setting a standard on the North Carolina High School Competency Test.** Paper presented at the spring meeting of the North Carolina Association for Research in Education, Chapel Hill, 1978.

Schoon, C. G., Gullion, C. M., & Ferrara,, P. Bayesian statistics, credentialing examinations, and the determination of passing points. **Evaluation and the Health Professions**, 1979, 2, 181-201.

van der Linden, W. J. Decision models for use with criterion-referenced tests. **Applied Psychological Measurement**, 1980, 4, 469-492.

Shepard, L. A. Setting performance standards. **A Guide to Criterion-Referenced Test Construction**, Edited by Ronald A. Berk. Baltimore, Md.: Johns Hopkins University Press, 1980, Chapter 7, 169-198.